Two Strands of Field Experiments in Economics : A Historical-Methodological Analysis

Nagatsu, Michiru

2020-01

Nagatsu , M & Favereau , J 2020 , ' Two Strands of Field Experiments in Economics : A Historical-Methodological Analysis ' , Philosophy of the Social Sciences , vol. 50 , no. 1 , pp. 45-77 . https://doi.org/10.1177/0048393119890393

http://hdl.handle.net/10138/313942
https://doi.org/10.1177/0048393119890393

acceptedVersion

Philosophy of the Social Sciences

Two strands of field experiments in economics: A historical-methodological analysis

Journal:	*Philosophy of the Social Sciences*
Manuscript ID	POSS-19-0059.R1
Manuscript Type:	Original Manuscript
Keywords:	experimental economics, field experiments, randomized experiments, internal and external validity, exhibits, history of economics, Development Economics, methodology of economics

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

Nagatsu_Favereau_Final_LaTeX2e class file for SAGE Publications.zip

Two strands of field experiments in economics: A historical-methodological analysis

August 30, 2019

Abstract

While the history and methodology of laboratory experiments in economics have been extensively studied by philosophers, those of field experiments have not attracted much attention until recently. What is the historical context in which field experiments have been advocated? And what are the methodological rationales for conducting experiments in the field as opposed to in the lab? This paper addresses these questions by combining historical and methodological perspectives. In terms of history, we show that the movement toward field experiments in economics has two distinct roots. One is the general orientation of medical and social sciences to evidence-based policy evaluation, which gave rise to *randomized* field experiments in economics (e.g., behavioral public policy, poverty alleviation policy, etc.) The other is an awareness of several methodological limitations of lab experiments in economics, which required practitioners to get out of the lab and into the field. In these senses, the movement is a consequence of influences from both outside and inside economics: the general evidence-based trend in policy science and an internal methodological development of experimental economics. In terms of methodology, we show that these two roots resulted in two somewhat different notions of "external validity" as methodological rationales of field experiment. Finally, we suggest that analysis of experiments as *exhibits* highlight a methodological strategy in which both strands methodologically complement each other.

Keywords: Experimental economics, Field Experiments, Randomized Experiments, Methodology of Economics, History of Economics, Internal and External Validity, Exhibits

JEL codes: B210, B410, C930, C900

1 Introduction

The experimental turn in economics in the last 50 years has invited substantial historical and methodological analyses (Guala, 2005; Bardsley et al., 2010; Heukelom, 2014; Svorenčík and Maas, 2015).[1] Many of these analyses focus on describing and justifying how economics, a discipline once regarded as non-experimental, has established a canon of laboratory experimental methodology despite skepticism from both within and outside the profession.

In contrast, the use of *field* experiments, an increasingly popular trend in economics, has just started to be reflected on methodologically and historically. While the existing discussions on this topic by practitioners (Harrison and List, 2004; Levitt and List, 2009; DellaVigna, 2009; Duflo, 2006; Charness et al., 2013) suggest the involvement of heterogeneous traditions such as behavioral and experimental economics, development economics, education economics and political economy, much philosophical attention has been narrowly focused on the methodological problems of randomization. This state of the art makes it difficult to understand the significance of field experiments in economics and their historical and methodological complexity.

The aim of this paper is to offer a historically-informed methodological analysis of field experiments in economics. We argue that once a basic but crucial distinction between two historical strands of field experiments in economics is made, their respective methodological advantages and weaknesses, and the ways in which they complement each other, will become clear. The first strand, which we call the

[1] This turn is part of the broader applied or empirical turn that economics have taken place in the last 50 years. For an analysis of these broader turns see **?**. In this paper we focus on the experimental turn of such broader turns.

lab-in-the-field experiments (LFEs) strand, extends the laboratory experimental economics to address the *artificiality* of traditional lab experiments. The second strand, which we call the randomized field experiments (RFEs) strand, applies randomized controlled trials, widely and increasingly-used in the social sciences on economic problems, to provide 'evidence' of 'what works' for policy. In this strand, the key methodological challenge concerns the *generalizability* of results from RFEs for policy. Although both artificiality and generalizability concerns are often framed as the problem of the external validity of extrapolation from experimental findings, this is misleading since their underlying methodological concerns are distinct. After making this distinction clear, we propose that the exhibit analysis (Sugden, 2005, 2008; Bardsley et al., 2010) highlights a useful methodological strategy to exploit each strands' methodological complementarity.

The paper proceeds as follows: Section 2 characterizes the two historical strands of field experiments in economics. Section 3 outlines the corresponding methodological rationales of each strands, and how they relate to the notion of external validity discussed in the methodological literature. Section 4 offers a strategy to exploit methodological complementarity of both strands of field experiments with an illustrative case. Section 5 concludes.

2 Two historical origins of field experiments in economics

In this section, we show that field experiments in economics have two distinct origins. One stems directly from the advancement of lab experiments in economics at the turn of the century (2.1). This strand gave rise to the lab-in-the-field experiments (LFEs) as a way to overcome the limitations of laboratory experiments: their artificiality relative to more naturalistic real-world settings. The second strand, the randomized field experiments (RFEs), originates from social field experiments in the second half of the 20th century, which have been later adopted to evaluate smaller-scale economic policy interventions (2.2).

Our basic distinction is not purely historical but motivated by the methodological notion of *controlled variation*. Controlled variation, or simply control, is a hallmark of experimental methods, which enables experimenters to establish a causal link between the putative cause and its putative effect in an experiment (Guala, 2005, 67). While the LFEs have their roots in the efforts of experimental economists to *directly* control variation, RFEs originate from the attempt to *indirectly* control variation by random assignment of subjects into treatment(s) and a control group. As we will see, these distinct strategies to achieve control in experiments give rise to distinct traditions and methodological concerns.

2.1 The first strand: Field experiments as an extension of lab economic experiments

As Guala (2008a) points out, experimental economics has its origins in many disciplines such as mathematics, philosophy, psychology, sociology and political science, in addition to economics, which makes it difficult for us to tell its history comprehensively. While there are different ways to tell this history (e.g. Roth, 1995), it is useful for our purposes to focus on the methodological notion of *control*. The first historical strand we identify originates in the attempt to increase control in the economics laboratory. The driver of this strand was to tighten control in order to secure causal inferences (2.1.1). At the same time, however, the lab creates an artificial space to secure control of variables of interest. This motivates a call to relax such control in order to decrease the artificiality of the lab. This historical strand, the lab-in-the-field experiments (LFEs), is defined by this move from tightening to relaxing control. However, LFEs stay close to the idea of direct control, as the way this relaxation takes place is systematic (2.1.2).

2.1.1 Tightening the control: The establishement of experimental economics in the lab (1950-2000)

According to the orthodox account (Roth, 1995), three historical origins can be identified in laboratory experimental economics. The first can be traced to the psychologist Thurstone's experiment in Chicago, which tried to measure individuals' indifference curves in the 1930s. This line has continued as experiments on individual judgement and decision-making under labels such as mathematical psychology and

behavioral decision research, forming the foundations for behavioral economics (see Heukelom, 2014). The second line starts with the experiments at RAND by the mathematicians Dresher and Flood in 1950, now known as the Prisoner's Dilemma game. The third line originates from the Harvard economist Chamberlin's classroom experiments to test market equilibrium predictions in the 1940s, leading to Vernon Smith's seminal contributions to market experiments.

A common element of experimental economics from these lines is its incremental and systematic effort to control for economically important variables. For example, Vernon Smith's (1976) Induced Value Theory proposes sufficient conditions for a microeconomic market experiment to achieve control for subjects' preferences over outcomes of their actions by using a medium of reward that (the experimenter knows) the subjects value in specific ways (Guala (2005, 232-233); Camerer (2003, 35)). Although psychologists' decision-theoretic experiments traditionally used hypothetical payments, the use of task-related incentives has also become more common through empirical and methodological debates with economists (see Hertwig and Ortmann, 2001, for the methodological differences between experimental economics and experimental psychology). As a result, requiring task-related monetary incentives has become part of the convention in experimental economics to control not only for preferences, but also for the features of the objects of choice, such as private-public goods and risky lotteries (Bardsley et al., 2010, 248-249). As this example of incentives shows, control has been a central interest of experimental economics, and the laboratory has been regarded as conducive to increasing control.

From this methodological vantage point, many of the early experimental studies

look more 'field-like' in the sense that they do not control for many variables. For example, Ledyard (1995, 124) criticizes a lack of control in one of the earliest public goods experiments by Bohm in 1972, in which subjects' willingness to pay, and belief about the group size, were unmeasured and uncontrolled. Another example is Chamberlin's (1948) market experiment, which allowed his students to move around freely in the classroom and negotiate bilaterally with one another without any predefined rules. These 'field-like' features have been gradually eliminated from experimental economics practices.

2.1.2 From the lab to the field: Relaxing control in a controlled way (2000-)

Since the late 1990s and 2000s, there has been a movement to conduct economic experiments "in the field" as opposed to the lab (Harrison and List, 2004). The early work of John List, one of the central figures in this movement, and collaborators (List and Shogren, 1998; List and Lucking-Reiley, 2000) aimed at studying methodological questions, such as whether or not the calibration of hypothetical and actual bidding is specific to goods and contexts. A variety of goods is needed to answer such questions, so they conducted a field auction experiment with actual sports card (baseball and football) collectors at a sports card show in Denver, CO in December 1995. As evident in this early methodological interest, the field experiments movement is a natural extension of lab economic experiments to address questions originating in the lab.

In addition to an inability to address such methodological questions, some of

the limitations of lab experiments have been recognized by the practitioners. One worry has been that an effort to standardize and control everything tightly in the lab might lead to a result that is not relevant to or informative about real world policy-making. Even worse, such artificial experiments may result in the unintended loss of control over relevant variables. For example, the unnatural or artificial lab environment might make the subjects confused about what the task is about, which means that the experimenter loses control over the subjects' framing of or beliefs about the game (Harrison and List, 2004). Therefore, the key slogan of experimental economists in conducting field experiments is that one should "lose control in a controlled way" (Bolton and Ockenfels, 2012, 671). Levitt and List (2009, 2) offer a useful characterization:

> This [...] movement approaches field experiments by taking the tight controls of the lab to the field. In doing so, the analyst bridges laboratory and naturally-occurring data by systematically relaxing the controls inherent in a laboratory experiment.

The key phrase here is "systematically relaxing the controls". Field experimentalists' responses to the problem of the artificial laboratory and the difficulties of interpreting and extrapolating the results to outside the laboratory wall have two components, namely systematicity and controls. We discuss them in turn.

Systematicity refers to how different experiments are designed, which allows the adding of one or more extra components that represent the natural settings in which economic decisions are made. Table 1 shows this systematic character of field experiments by comparing the influential taxonomy by Harrison and List (2004) and

a more recent alternative by Charness et al. (2013). While details of terminology differ, both taxonomies are defined in terms of the systematic ways in which the modifications are made to the lab experiments (see Table 1). The table makes explicit an important methodological characteristic of field experiments in this strand, namely the systematic way in which experimental variables are varied.

Taxonomies		**Variables**			
Charness et al.	**Harrison&List**	**subject**	**referent**	**location/situation**	**awareness**
Conventional lab exp.		students	imposed	artefactual	yes
2*Extra-lab exp.	Artefactual field exp.	nonstandard	imposed	artefactual	yes
	Framed field exp.	nonstandard	field	artefactual	yes
2*Field exp.	n/a	nonstandard	field	natural	yes
	Natural field exp.	nonstandard	field	natural	no

Table 1: Taxonomies of field experiments by Charness et al. (2013) and Harrison&List (2004)

For example, if you observe a fair offer in the Ultimatum game in the conventional lab experiment, but observe a less fair offer in the experiment with non-student subjects, you can infer that this difference is likely to be due to characteristics of subject pools (Henrich, 2004). Similarly, if you observe differences in how people choose in the Diner's dilemma game in both artefactual and framed field experiments, then you can infer that the difference is due to the nature of the task or referents, e.g., a dish in a restaurant versus an abstract experimental currency on a computer screen (Gneezy et al., 2004). In contrast, if you see no difference in the results between these experiments, then you become more confident that those differences do not matter.[2] The main strength of this approach is that you can attribute a cause of difference in observations to one difference maker. That is, you can make a stronger inference about what causes what by changing one variable of interest while

[2]For more examples, see Gerber and Green (2012, 10, footnote 20).

holding the others constant, which Guala (2005, 67) calls *controlled variation* in the context of laboratory experiments. Although controlled variation in this context is far from perfect, since there are many other potentially relevant variables that are different across experiments, the idea behind comparing findings across these types of experiments is clearly analogous to controlled variation in the laboratory.

A famous example of this type of comparison is a labor markets experiment by Gneezy and List (2006), which is a (natural) field version of Fehr et al. (1993). The latter experiment was designed to test the theory of "gift exchange" originally proposed by Akerlof (1982), who argued that in labor markets employers offer wages higher than their reservation utility and employees in turn reciprocate with extra effort, which is distinct from the prediction of sub-game perfect equilibrium. Gneezy and List (2006) demonstrate substantial but short-lived reciprocal behavior of employees in natural one-time task job markets.

The relaxation of controls seems to contradict the idea of controlled variation, but this is not the case. Note that the notion of control is used in an ambiguous way, as in the phrase "losing control in a controlled way". The latter control refers to controlled variation, which is considered to define experimental science (e.g. Camerer, 2003). In contrast, the former control refers to artificial features of the lab experiments, which are undesirable either (i) because they lead to unintended loss of control (e.g. List, 2007; Bardsley, 2008) or (ii) because they make interpretation difficult by creating an artificial environment that does not correspond to naturally occurring economic settings in any straightforward way. In sum, although practitioners disagree about the exact taxonomy of experiments as well as the extent to which the field experiments

should be tied strictly to the lab experiments (e.g. Bardsley et al., 2010, 241), there is a common understanding of the importance of controlled variation and concern about the artificiality of lab experiments.

2.2 The Second Historical Strand: Field Experiments as a Tool for Evaluating Policies

Field experiments in economics are also directly inspired by large social scale public policy evaluation. Such inspiration defines the second strands of field experiments in economics. During the first part of the 20th century, field experiments were advocated as a useful tool to evaluate large scale public policy. Economists, more specifically labor economists, implemented several field experiments to assess such large scale policies (2.2.1). More recently, field experiments have been extensively conducted in economics in the last decade, in particular within development economics, to assess the impact of aid development programs (2.2.2.). Such evaluations are local rather than large-scale. However, they both emphasize the statistical rigor that the random allocation design of the field experiments allows. The random assignment is a particular and indirect method of achieving controlled variation, the emphasis on which defines the second strand of field experiments[3], in contrast to the first strand, in which the notion of control is not tied to particular methods such as random assignments of treatments.

[3]For a discussion about the control or balance over variables in randomized experiments, see Heckman (1992), Heckman et al. (1998), Heckman and Smith (1995), Heckman et al. (1997).

2.2.1 Social field experiments in economics (1960-2000)

As stated above, many field experiments in economics are rooted in the history of social field experiments. The latter aim to assess public policies, primarily on a large scale (Levitt and List, 2009). Greenberg and Shroder (2004), who have provided the largest work on reporting all the social experiments that have been conducted so far, define a social field experiment as an experiment that aims to (1) assess the impact of (2) a political intervention by (3) collecting data on such political intervention. In order to properly collect data, the participants of the experiment are (4) randomly assigned into at least two groups. In one of these two groups the participants receive the political intervention that is being assessed and the other group do not. Such a design allows the comparison of the two groups and thus allows the impact of the intervention to be identified. The random assignment also provides another advantage: it removes the selection bias that confounds results. The evaluation of public policies is the focus of such social experiments. Indeed, the primary objective of these experiments is to "speak to policy-makers" ((Greenberg and Shroder, 2004, 4). Greenberg and Shroder's (2004) definition exclude many experiments conducted in economics which do not satisfy (2) and (3).

Ferber and Hirsch (1982) define field social experiments in a very similar manner to Greenberg and Shroder (2004) but clearly relate them to the domain of economics, suggesting that a social field experiment evaluates the impact of both economic and social political intervention. A social field experiment is defined by four characteristics in the definition, namely, (1) public funding; (2) a rigorous statistical design; (3) targeting human subjects; (4) assessment of the effect of a political intervention

on social and economic variables.[4]

Common characteristics in these two definitions are rigorous statistical design and the evaluation of political intervention, which enable proper assessment of the impact of public policy. Indeed, Ferber and Hirsch (1982) underline the fact that although other methodologies were available to evaluate public policy during the second half of the 20th century, they were not as reliable as field experiments, and especially in economics, because of the many unobservable variables. Thus, social field experiments appear to be a very powerful tool for economics because of their potential to control for unobservable variables. The major difference between the two definitions is that Ferber and Hirsch (1982), but not Greenberg and Shroder (2004), insist on public funding, highlighting the fact that such experiments are extremely costly because of their large scale interventions.

During the second half of the 20th century, labor economics constituted a major domain in economics where field experiments were implemented. During the Reagan administration in 1986, a vast job program was also evaluated through a field experiment, which was also randomized: the Job Training Partnership Act Title II-A. This program offered financial support and training to people facing important barriers on the job market (such as low qualifications). This experiment is the paradigm case used by Heckman et al. (1997) to question the power of social evaluations in contrast to structural econometrics. Since then, more than 235 social field experiments evaluating public policy programs have been conducted over the second half of the 20th

[4] "A publicly funded study that incorporates a rigorous statistical design and whose experimental aspects are applied over a period of time to one or more segments of a human population, *with the aim of evaluating the aggregate economic and social effects of the experimental treatments.*" (Ferber and Hirsch, 1982, 7, our emphasis)

century (Greenberg and Shroder, 2004) offer an overview of all these experiments).

2.2.2 Randomized field experiments for evidence-based policy (2000-)

During the turn of the century, field experiments began to be advocated not only as a useful tool for policy evaluation but as a methodologically rigorous approach to produce evidence for policy in general. As noted above, Greenberg and Shroder (2004) and Ferber and Hirsch (1982) defined one of the key characteristics of social field experiments as a random assignment of the experimental subjects into at least two groups. Such assignment enables control for both observable and unobservable variables, thereby offering a reliable evaluation of the impact of policies. During the 1990s this reliable evaluation was highlighted, following the trend first initiated within the movement of evidence-based medicine.

The random assignment (one of the main characteristics defining a randomized field experiment) allows for reliable results by removing key statistical biases. Specifically, an experiment needs to control for both observable and unobservable confounders to offer reliable results. Even carefully designed laboratory experiments (see 2.1) cannot control every variations, for the simple reason that the experimenters cannot foresee every possible variables that might affect the results. While the LFEs strand tries to achieve controlled variation through the laboratory methods such as Induced Value Theory, the RFEs strand adopts an indirect control through randomization as the central method. Random assignment of the participants to at least two groups removes or balances selection bias.[5] Biases concern the possibility that

[5]See again here the discussion of Heckman (1992) for example.

other factors than the one the experiment is testing (e.g. the treatment) would explain the experimental results (e.g. the unmeasured differences between the groups). In other words, randomization allows the experimenter to isolate only the effect of the treatment. Thus randomized field experiments possess a strong internal and statistical conclusions validity. Such strong internal validity gives RFEs' results the status of evidence,[6] leading, as we have seen in the previous section, to the so-called evidence-based movement.

In 1992, a seminal paper was published by a group of thirty researchers who labelled themselves the evidence-based medicine working group (Guyatt et al., 1992). The paper argued that advances in medicine have been substantial thanks to the recent massive use of randomized controlled trials (also using random assignment as the key aspect of the experimental design); however, such scientific advancement was not used in daily medical practice. The claim of the working group was that the practice of medicine should be based on evidence provided by randomized controlled trials. Randomized experiments are thus used and justified during this period for two concomitant reasons: (1) the fact that they produce evidence, (2) the fact that this evidence can directly help decisions-makers (here physicians).

At the beginning of the 21st century, the social sciences followed the same trend, using social field experiments to base policy decisions (instead of medical ones) on evidence. The two key characteristics defined by both Greenberg and Shroder (2004) and Ferber and Hirsch (1982), the random assignment and the policy evaluation aim, are still present, but the emphasis has shifted from the latter to the former. Tony

[6]for a discussion about such status see Cartrwight and Duflo REF

Blair was one of the first to popularize the idea of and the term "evidence-based policy" in 2000. The United States and Australia also played a major role in the development of evidence-based policy. As this trend shows, the evidence-based policy movement is prominent in the Anglophone part of the world, whereas its introduction was later and smaller in other parts of the world, for example in France, even though some randomized experiments were and are still being conducted. RFEs are extensively used in development economics in particular, since the creation of the Jameel Abdul Latif Poverty Action Lab (J-PAL) at the Massachusets Institute of Technology (MIT) in 2003 by Abhijit Banerjee, Esther Duflo and Sendhil Mullainathan. J-PAL, one of the main centers of evidence-based policy movement, aims to systematically promote the implementation of RFEs in order to provide evidence to policy-makers of developing countries about the efficacy of development programs.

On the one hand, field experiments were introduced in economics to overcome the limitations of lab experiments, by offering them more "natural" settings along different dimensions than in the lab. On the other hand, field experiments entered economics as a tool to assess policy interventions and then as the main tool to produce evidence to guide policy-makers. The history of field experiments in economics, we suggest, becomes clear once these two strands are clearly distinguished. These two historical strands also create two methodological rationales for distinct types of field experiments, lab-in-the-field experiments (LFEs) and randomized field experiments (RFEs). These two methodological rationales originate in two distinct approaches to controlled variation, a direct one for the LFEs and an indirect one for the RFEs.

3 Two faces of external validity

The principal methodological concern for experiments in general is that of their validity. Commonly, the validity of experiments is defined as twofold: internal and external. Internal validity refers to the validity of an experiment's result within the frame of that experiment. Internal validity thus refers to the level of control one experimental frame allows; the more control, the stronger the internal validity. In contrast, external validity relates to the validity of the experiment's result outside the frame of the experiment, and it has been discussed more extensively by philosophers and practitioners alike.

Our aim in this section is to illuminate this methodological debate concerning the external validity of field experiments in economics by drawing on the distinction between the two field experiment strands we have defined in the last section. Specifically, we argue that the two strands are concerned with two distinct types of external validity issues: what is at stake concerning the external validity for the RFE strand is the possibility to *generalize* the experiment's result to another setting, population, context, etc. This challenge comes from the indirect way in which RFEs achieve control. We call this the issue of *generalizability* (3.1); in contrast, the LFEs strand, with its strong connection to the lab and thus a high degree of direct control over variables, creates an experimental frame that can be characterized as *artificial* relative to the naturally-occurring decision-making contexts which the experiment intends to inform. This concerns the issue of *artificiality*. Artificiality in this sense diminishes the external validity of the results of lab experiments, but for distinct reasons (3.2).

3.1 Validity concepts in experimental social science

As we have discussed above, the notion of validity is commonly used as a methodological criterion to evaluate experiments. Internal and external validity concepts were first developed during the 19th century in psychology as the "accuracy of a [psychological] test" (Heukelom, 2011). During the 1950s they came to mean the accuracy of the experiments conducted in psychology. In the middle of the 1950s the American Psychological Association (APA) aimed to specify the methodological standard for psychology and produced a textbook with methodological recommendations for doing research in psychology. Donald Campbell (1957) opposed those recommendations, which led to several debates between Campbell and the APA that were concluded by a settled definition of internal and external validity in psychology.[7] The resulting definitions were immediately adopted within the RFE strand. The historical and methodological continuity between psychology and the RFE strand might explain this immediate conceptual adoption. Building on the seminal definition, Shadish et al. (2002) distinguish four types of validity: statistical conclusions validity, internal validity, construct validity, and external validity.

1. Statistical conclusions validity refers to the existence of a statistical variation and its degree; it asks "whether the presumed cause and effect covary and how strongly they covary" (Shadish et al., 2002, 42).

2. Internal validity shows that the observed statistical variation is a causal relation

[7] "The discussions of validity in the Technical Recommendations and by Campbell between the early 1950s and early 1960s defined the experimental vocabulary in the psychology for subsequent decades, up into the twenty first century." (Heukelom, 2009, 6)

and not a correlation and that this statistical variation is due only to the treatment that is being assessed through the experiment.[8]

3. Construct validity concerns the validity of measurement, in particular the labeling of the operation of an experiment. In order to study a phenomenon it is necessary to link its operationalization and the label; for example, to study intelligence, it is necessary to operationalize the notion using a certain psychological test. Such a test of intelligence is said to be constructively valid if it really measures the intelligence that we are interested in, and not something else: "construct validity involves making inferences from the sampling particular of a study to the higher-order constructs they represent" (Shadish et al., 2002, 65).

4. Finally, external validity is about the generalization of a causal inference from the context of the experiment to another one, and-or from the population of the experiment to another one, and-or from the treatment tested in the experiment to another one.

These four types of validity are deeply linked to each other, in particular (1) and (2), and (3) and (4). (1) Statistical conclusion validity and (2) internal validity are both concerned with the statistical variation between the treatment that is tested through the experiment and the experimental result. They concern two different

[8] "[W]e use the term internal validity to refer to inferences about whether observed covariation between A and B reflects causal relationship from A to B in the form in which the variables were manipulated or measured. To support such an inference, the researcher must show that A preceded B in time, that A covaries with B (already covered under statistical conclusions validity) and that no other explanations for the relation are plausible." (Shadish et al., 2002, 53)

steps of one process: the former is about the validity of the inference that a statistical variation exists, and the latter concerns the validity of the inference that a variation is a causal one. Internal validity thus presupposes statistical conclusion validity. Similarly, (3) construct validity and (4) external validity are deeply linked. While the former concerns the validity of the degree of generalization from particular research operation *to* higher-order constructs, the latter concerns the validity of the degree of replication *across* different constructs and populations. To use a spatial metaphor, the former regards *vertical* extrapolation, the latter *horizontal*. The term "external validity" in the literature of experimental economics does not distinguish between the two, probably due to the general disciplinary distance between economics and psychology.[9] But the distinction is relevant to understanding the methodological challenges for both strands of field experiments, as we shall see below.

3.2 External validity as the problem of generalizability of randomized field experiments

As we saw in the previous section, while the internal validity of RFEs seems to be strong, their external validity has been extensively criticized from several perspectives, includign the philosophy of science (Cartwright, 2007, 2009, 2010; Cartwright and Hardie, 2012; Teira, 2013; Teira and Reiss, 2013; Davis, 2013), development economics (Deaton, 2010; Ravallion, 2009; Rodrik, 2008; Acemoglu, 2010; Basu, 2014; Barrett and Carter, 2010), experimental economics (Harrison, 2011, 2013), labor

[9]In contrasting the two disciplines, Ross (2019) points out that "[t]he language of 'constructs' is foreign to economists; referential terms in their models are taken to be directly isomorphic to real objects and processes."

economics and econometrics (Heckman, 1992; Heckman and Smith, 1995; Heckman et al., 1997, 1998; Leamer, 2010). Even though these criticisms target different methodological, practical and technical aspects of RFEs, they all somehow concentrate on the fact that RFEs produce a result termed "black-box test of 'what works"' (Deaton, 2010, 451). In other words, RFEs' results would not illuminate the mechanisms behind such results, making their possible evidential use difficult or even impossible. Indeed, both observables and unobservables that could offer explanations about *why* a treatment is working are masked by the random allocation of the treatments between the groups. For instance, an experimental result highlighting the efficacy of deworming in a specific region in Kenya (see Miguel and Kremer, 2004) to reduce students' absenteeism might turn out otherwise in India. Indeed, in India students might not suffer from worms but rather from anemia (Bobonis et al., 2006), therefore iron would be effective while deworming would be not. Not knowing the mechanisms behind the experimental result (e.g., efficacy of deworming) does not tell us if such results will hold up in another context. Therefore, it appears difficult to *generalize* from the RFE results, and thus to produce policy recommendations beyond particular experimental set-ups. In order, to favor such generalizability, J-PAL researchers, such as Abhijit Banerjee and Esther Duflo, propose to concentrate on "structured speculative judgements" (Banerjee et al., 2017), or to combine machine learning and RFEs Chernozhukov et al. (2018). However, such propositions are still at an early stage of development.

3.3 External validity as the problem of artificiality of lab experiments

As Heukelom (2011) points out, the experimental economists who established the lab tradition did not refer to the term "external validity" until the 1990s. Heukelom (2011) suggests that experimental economists' reluctance to adopt the methodological distinction between internal and external validity is due to their fear that it would have pushed experimenters in economics to distinguish between an "inside" and an "outside" world.[10] As Heukelom (2011) points out, a series of methodological analyses overcame this reluctance and made it a key methodological concept for laboratory experiments in economics (Guala, 1999, 2003, 2005; Guala and Mittone, 2005). Indeed, Guala explicitly defines internal and external validity in terms of artificiality, highlighting the distinction between an "inside" and an "outside" world.

Experimental economists have acknowledged and analyzed this gap between the lab and the naturally occurring economic decision-making set-up as *artificiality*. In particular, Bardsley et al. (2010) define four types of artificiality: the artificiality of (i) isolation, (ii) omission, (iii) contamination, and (iv) alteration. (Bardsley et al., 2010, 215). The development of the lab-in-the-field experiments (LFEs) strand can be seen as a response to these artificiality concerns.

Regarding the artificiality of isolation and omission, this is an inevitable consequence of experimental control to study one putative cause at a time, in contrast

[10] "The reason that [Vernon] Smith and other experimental economists in the 1970's such as Plott were reluctant to adopt validity as understood by the psychologists was that the psychologists' way of validity risked creating a division between an inside world of the laboratory and an outside 'real' world of the economy and its actors" (Heukelom, 2011, 20)

to the natural environment in which multiple causes are at work at any given time. As we have seen, the solution proposed by LFEs is to study a putative cause while reducing isolation and omission in a controlled way. For example, lab phenomena such as gift-exchange or the endowment effect were first studied in the conventional labs, and then in the LFEs by varying referent goods, locations and situations closer to the field counterparts we are interested in.

The artificiality of contamination refers to the possibility that the participants' very awareness of being studied will introduce new causes, such as a wish to confirm to or disrupt the perceived goal of the experimenters.[11] What Harrison and List (2004) call natural field experiments address this problem by making participants unaware of the fact that they are being studied.

While these types of artificiality are methodologically well-addressed, at least in principle, the less discussed artificiality of alteration may pose a bigger challenge to the LFEs as it refers to the gap between the lab and the target phenomenon of interest, which cannot be filled in an incremental way. In other words, going to the field with standardized experiments does not solve this problem, but already assumes that the alteration does not happen. The alternation refers to the possibility that the experimental setup changes the nature of the phenomenon that the experiment is designed to study. For example, do behavioral experiments to study risk, social and time preferences represent naturally occurring situations in which people make relevant risky, interpersonal, and intertemporal choices? In other words, do risk, social and time preferences measured in the experiments measure the same preferences

[11] See Jimenez-Buedo and Guala (2016) for a detailed discussion of contamination as reactivity.

that operate in the wild? We will discuss this challenge in detail with examples in the next section.

The external validity problem thus construed as the artificiality of alteration is, we argue, distinct from the external validity generalizability problem that the RFEs strand faces. To use Cook et al.'s (2002) distinction that we introduced earlier, artificiality by alteration concerns (3) construct validity while generalizability concerns (4) external validity. Although both concerns the validity of extrapolative inferences, the grounds for extrapolation are different. Construct validity is grounded in the assumption that the psychological or economic construct of beliefs or preferences, operationalized by a particular elicitation procedure of an experiment and conceptualized or identified by a standard theoretical model, represents what it intends to represent. External validity, in contrast, is grounded in a more specific assumption about the comparison between the experimental model and the target phenomenon. Although in this paper we use the term external validity in a broader sense to refer to both generalizability and artificiality problems, thus departing from Shadish et al. (2002), the distinction is important for the methodology of field experiments in economics.

In this section, we have argued that the RFEs and LFEs strands face two distinct challenges, generalizability and artificiality, which are both discussed under the term external validity in the literature in experimental economics. We suspect that this is part of the reason why philosophers contest the very usefulness of the notion of external validity (Jiménez-Buedo, 2011; Reiss, 2018). Our proposal has been to make a methodologically relevant historical distinction explicit, before giving up on external

validity on conceptual grounds. In the next section, we propose a methodological strategy to address both concerns, while making the two strands complement each other in contemporary practices of field experiments in economics.

4 Integrating the two strands

Despite their distinct historical origins, the two strands of field experiments are gradually becoming mixed in practice, resulting in hybrid field experiments (Viceisza, 2016). What are the affordances of such hybrid experiments? In particular, how do they address the two distinct methodological issues defined in the previous section? We argue that there is some methodological complementarity between RFEs and LFEs, in particular the latter informing the data-generating process of the former. However, although LFEs become less artificial by being embedded in RFEs, the former's construct validity remains a problem.

To provide a methodological strategy in which both strands play complementary roles, we draw on the notion of *exhibits* introduced by Robert Sugden (4.1). We then show how this strategy combines both strands to address respective external validity concerns and inform both theory and policy (4.2). Finally, we illustrate our point by examining a recent field experimental study (4.3).

4.1 Three functions of economic experiments as exhibits

Roth (1995) distinguishes three functions of economic experiments, namely *informing theory*, *searching for facts*, and *informing policy*. In terms of this three-part catego-

rization of economic experiments, the two strands of field experimental traditions—the 'lab-in-the-field' experiments and randomized field experiments—are strongly associated with the first and the last functions, respectively. But we propose that a key to understanding the different external validity concerns of both strands in a coherent framework and realizing their complementary roles is to pay more attention to the second, fact-searching function. For this purpose, we draw on the notion of *exhibits* developed by Sugden (2005, 2008) and his collaborators (Bardsley et al., 2010).[12] An exhibit is "a replicable experimental design that reliably produces some interesting result" (Bardsley et al., 2010, 156). Examples include famous anomalies in behavioral experiments, such as Allais and Ellsberg paradoxes in judgements and decision-making under uncertainty. There are three things to highlight in this definition of an exhibit.

First, an exhibit as an experiment produces a phenomenon, which is distinct from what occurs naturally in the world.[13] In other words, the produced phenomenon is an experimental artefact. Daniel Kahneman calls such an experimental invention as an act of 'bottling phenomena', crediting the social psychologist Lee Ross for coining the expression (Andersson and Holm, 2002, 45).

Second, an exhibit has to produce a phenomenon in a reliable and replicable manner. That is, the experimental design is expected to produce the same phenomenon, within a range of design variables. Replicability makes the bottled phenomenon dis-

[12]Several authors have developed related notions such as paradigmatic experiments (Guala, 2008b; Guala and Mittone, 2010) building on exhibits. For our purposes, however, we mainly use Sugden's original term.

[13]In the following literature some define an exhibit as a phenomenon. We follow Sugden's original definition and distinguish the exhibit as an experiment and the phenomenon it produces.

tinct from a 'mere artifact', which is thought to be caused by some idiosyncratic features of the experiment and therefore unworthy of further investigation. In contrast, the reliability of an exhibit to produce the same phenomenon motivates experimenters to vary the original design systematically to investigate its regularity and mechanisms.

Third, the resulting phenomenon is characterized as 'surprising' or 'interesting' to the relevant research community. Often the surprise comes from the fact that an exhibit constitutes an anomaly to the standard theory, as in the case of Allais paradox violating the independence axiom of expected utility theory, or the fair offer in the ultimatum game contradicting the sub-game perfect equilibrium prediction of minimum offer. However, a phenomenon can be interesting or surprising without clearly contradicting the prediction. For example, in finitely repeated public goods games with the voluntary contribution mechanism, contributions typically start off around 50% in the initial round, and then steadily decline toward the last round (Ledyard, 1995). This pattern seems to confirm the free-riding hypothesis (subjects are learning how to play rationally), or disconfirm it (the contribution usually does not reach zero even in the final round), depending on what one sees in the phenomenon. Either way, the phenomenon has interested a generation of experimental economists, who have conducted numerous experimental variations and offered different explanations. Thus, although surprise or interest sounds subjective, it is defined against some shared or contested background knowledge in the research community. A phenomenon may also simply be 'unexplained by' a received theory, in the sense that the received theory does not take into account its alleged causal mechanism,

in which case the phenomenon neither confirms nor disconfirm the theory. The fact that the initial contributions in public goods games are typically around 50% is an example of such phenomena.

What are exhibits good for, then? Sugden proposes that, in addition to challenging existing theories and helping in the construction of new ones (i.e., theory-testing and building), exhibits serve at least three distinct functions in empirical investigations.

First, they serve as a "public arena in which empirical insights can be presented and discussed, structured so that the more promising insights tend to have more prominence" (Sugden, 2005, 298). This social-epistemic function has two mechanisms. First, an exhibit provides an inter-subjective reference phenomenon that other scientists can verify or challenge, and that the subsequent research can refer to (Bardsley et al., 2010, 163), which we call coordination. Second, Sugden (2005) also conjectures that an exhibit may attract many rival explanations for the phenomenon when multiple causal mechanisms are at work *in the same direction* to create the phenomenon (attention). In other words, while there are many ways to coordinate research efforts, exhibits may tend to facilitate coordination on causally important phenomena. On this account, a prominent exhibit will help produce and sharpen alternative explanations and enrich our causal understanding of that relevant phenomenon, without necessarily resulting in a replacement of the standard theory by some unified theory that wins the competition. Still, this process facilitates empirical progress in two ways. On the one hand, an exhibit's ability to produce a robust regularity can be examined through a range of changes in experimental design (robustness

check). On the other hand, an exhibit can be further broken down to finer-grained exhibits that demonstrate regularities (decomposition) (Bardsley et al., 2010).

A second function of an exhibit is to serve as a *construct*, or an experimental measure or representation of some feature of real-world behavior (Sugden, 2008, see also (Guala, 2008b; Guala and Mittone, 2010)). Shadish et al.'s (2002) discussion of construct validity (see Section 3.1 above) is useful in explicating this function. When an exhibit produces a valid construct, the latter represents some higher-order, more *abstract* disposition of a given population. Sugden's example is 'trust' in the trust game.[14] In the trust game, a varying level of 'trust' is measured as the ratio of the endowment transferred from the Truster to the Trustee. The level of 'trust' is expected to vary across populations, capturing their different levels of trust. In other words, valid constructs permit *vertical* extrapolation to the more abstract target phenomena, rather than *horizontal* extrapolation to other concrete situations. How do we increase the validity of a construct? Guala (2008b) emphasizes the importance of the *standardization* of experimental design, or the establishment of what he calls 'paradigmatic experiments'. Clearly, however, standardization as such is not sufficient because assuming so would imply that each standardized operation has its own construct, which undermines the attempt to let the procedure measure a higher-order construct. As we discussed, this worry, the artificiality of alteration, cannot be addressed by merely moving the standardized experiments to the field

[14]The trust game (Berg et al., 1995) is a two-person sequential game in which the first mover (Truster) will decide the proportion of transfer of the initial endowment, which will be multiplied by a certain factor; the second mover (Trustee) then decides the proportion to return. Although trust and reciprocal cash transfer can benefit both players, since the Trustee can maximize his income by not returning anything, and since the Truster knows this, the sub-game perfect equilibrium is no transfer, or no trust.

(LFEs). What is additionally needed is the more or less *convergence* of the measures from these experiments, as well as between these measures and other indicators of the construct coming from various non-experimental data on a given population, such as survey data about trust in the government, prevalence of trustful behavior in the field, etc. In this sense, LFEs are not a sufficient but necessary step toward building a valid construct.

Finally, exhibits serve as analogous surrogates for 'real-world' phenomena to be compared, in particular behavior in non-experimental settings. This means that by studying the causal mechanism of the exhibit we aim to learn about that of the analogous real-world behavior in question. Sugden's example is a lab coordination game, to which some observable real-world regularity, namely the tendency for trade deals to be struck at round numbers (Schelling, 1960, 67), is compared. To reason that the lab experiment and the real world regularity share some relevant mechanism is a type of inductive reasoning called analogical reasoning (see Guala, 2010; Steel, 2010). Although Sugden is very cautious about identifying the mechanism of an exhibit with that of a target real-world phenomenon, he does not exclude the possibility of using exhibits as "explanatory devices in their own right" (Sugden, 2005, 298) in this way. Since analogical reasoning is fallible, several methodological approaches have been proposed to make it successfully, the most prominent one being comparative process-tracing (Steel, 2008).[15] In what follows, we highlight complementary roles of LFEs and RFEs, focusing on which functions each type of experiments serve.

[15]See Khosrowi (2019) as a critical evaluation of comparative process-tracing as a research strategy in econometrics.

4.2 Combining RFEs and LFEs as exhibits

How can RFEs and lab-in-the-field experiments (LFEs) be construed and combined as exhibits? In this subsection we present a schematic way to do so, which will be illustrated with an actual experimental study in the following sub-section.

Table 2 summarizes the three functions of exhibits we discussed above in column, applied to RFEs and LFEs in row.

Functions of exhibits	Coordinate (Attention)	Construct (Abstraction)	Compare (Analogue)
Randomized field experiments (RFEs)	(1)	-	(3)
Lab-in-the-field experiments (LFEs)	-	(2)	-

Table 2: Functions of field experiments as exhibits: Circled numbers indicate the step-wise hybrid use of RFEs and LFEs that we advocate.

First, we discuss whether and how RFEs perform the three functions as exhibits, in comparison to lab experiments. Then, we argue that, when combined, RFEs and lab experiments, in particular LFEs, complement each other in terms of these functions.

First of all, RFEs, just like lab experiments, fulfill the social-epistemic function as research coordination devices ((1) in Table 2). That is, some RFEs can attract research efforts by their prominence by virtue of their surprising character (against some theoretical background), replicability and accommodation of multiple causal explanations. For example, the exhibit that the poor do not invest enough in cheap preventive health products, whereas they should if they follow rational choice models (Dupas, 2014) illustrates such a surprise. When such behavior is observed not only

for bednets, but also for other products such as deworming pills and chlorine, the exhibit is more replicable, suggesting some regularity in health-related behavior in general. And when the exhibit invites multiple explanations (present-bias, social pressures from family and relatives, credit constraints, etc.), its prominence increases. Another example is an experiment that shows that behavioral intervention that the received theory considers to be of no relevance turns out to have a significant impact on behavior, which also invites more experiments and explanations.

Concerning the second function of exhibits, namely to serve as a construct, RFEs are not suited for this purpose. This is because they are not as standardized as lab experiments, which is a necessary feature for them to function as portable measurement devices of particular constructs (Guala, 2008b). RFEs are primarily designed to measure the impact of interventions, not the level of trust or altruism in a given population. Lab-in-the-field experiments, in contrast, are intended to function this way, as their standardized experimental designs are implementable in different locations with different populations (② in Table 2). When a LFE measures a valid construct, it can be used to tell us whether and to what extent a given population is heterogeneous in terms of the level of the construct it measures. As we will see below, this information facilitates the use of RFEs as surrogates for counterfactual scenarios of policy interventions.

Regarding the third function of serving as an analogue to be directly compared to the real world target system, RFEs are intended to serve as such surrogates, as they are intended to demonstrate what works in the real policy contexts (③ in Table 2). However, as we have seen in the last section the main concern with RFEs is their

generalizability. In particular, it is difficult to soundly infer that what happens in an RFE is going to happen in an analogous counterfactual policy scenario, in which a similar intervention in another sample, time, scale, or location is implemented. This use of RFEs as surrogates is more demanding than the case Sugden considers, in which a lab experiment is compared to some observable real-world regularity. Unlike this use of a lab experiment as an explanatory device, a RFE is intended to function as a simulation device that tells us what would happen if a particular policy intervention would be implemented in a particular setting. This demands more careful and detailed confirmation of the relevant analogy between the RFE and its implementation in the intended context. In the next sub-section, we will illustrate how the use of LFEs as measures of constructs will help such analogical reasoning.

4.3 Illustrations

Our example comes from studies on health-related purchasing/saving behavior in developing countries. As underlined above, RFEs have suggested that the poor do not invest in preventive health products although they are relatively affordable; instead, they often end up spending a substantial amount of money on curative products, while incurring the cost of diseases (see Dupas, 2011, for a review). RFEs on bednets against malaria illustrates this more in detail. First, the demand for bed-nets drops by 80% as soon as they cost some money (Cohen and Dupas, 2010, 20). Second, individuals who benefited from bed-nets for free in an initial period are more willing to pay for bed-nets (Dupas, 2014). Third, the poor do not invest in bed-nets, whereas they are intensively used when given out for free, which Dupas

(2009, 230) calls a "puzzling fact".

We propose that this puzzling fact can be construed as an exhibit in the first sense. Indeed, it highlights a surprising fact contradictory to standard rational choice theory. It would be rational for individuals to invest in cheap and effective preventive health care (such as bed-nets) rather than to spend more on expensive curative health products. Such a fact is not only surprising but also regularly observed in many RFEs (Dupas, 2011). The fact that similar RFEs show similar patterns with other preventive health products such as deworming pills and chlorine tablets to purify water in other regions point to the regularity of the exhibit. By producing a surprising experimental phenomenon, these RFEs help coordinate research attention and efforts.

However, RFEs alone are unable to explain such results as they are like a black-box (e.g. Deaton, 2010). What explains the under-investment in products with apparent low cost and high expected returns? Mahajan and Tarozzi (2012) provide an excellent case in which this question was addressed in a large-scale RFE that evaluated alternative micro financial tools to provide insecticide treated nets (ITNs), and how they impact health and socio-economic outcomes of potential users in India.[16]

[16] The study is part of the broader project published as Tarozzi et al. (2014). We are focusing on this study for the following reasons: first, its topic, the ITNs use, is the same as Dupas's RFEs that gave rise to the more general phenomenon that we have been discussing, namely the poor's under-investment in preventive health products; second, we consider the methodological standard of the study to be high, even compared to other published studies. Although it is still a working paper, it has already been cited 55 times, and its larger study, the RFE part, (Tarozzi et al., 2014) has been published in *American Economic Review* and cited 162 times (both on Google Scholar, checked on 27th August 2019), suggesting that the study is also well-received in the research community. In general, in economics many influential and important texts are published in respected working paper series, followed by publications in journals a few years later. See Viceisza (2016) for a more systematic review of more examples of hybrid experiments, many of which are working papers as of 2016.

The motivation of the study was to find an effective micro-credit for the poor to buy ITNs as an alternative to free distribution. Given that health organizations and local governments also have budget constraints, free distribution of preventive health products is not a sustainable policy option. Moreover, free distribution will target the most credit-constrained households more effectively if other households share or fully pay the cost of the products using existing affordable micro-credit schemes. In order to devise such effective schemes, however, we need to understand why some households choose particular micro loan scheme while others choose another, and yet others do not purchase ITNs at all.

Mahajan and Tarozzi (2012) addressed this question by conducting a LFE inside the RFE. In particular, the lab-in-the-field part of the experiment focused on what explains the take-up rates of different micro-loan schemes to purchase ITNs. In the RFE they gave the households two alternative micro-loan offers through a local micro lending association, one with the service of future re-treatments of the net, and the other without. The former contract has a feature of a commitment device, binding the buyer to the payment of the future re-treatments, but with lower cost than when the buyer later decides to pay for re-treatments. The households also had the option to not purchase any net. So within this treatment group, the observed choices were (i) purchase of a net with a commitment loan; (ii) purchase of a net without a commitment loan; and (iii) no purchase.[17] The question that the RFE alone cannot answer is: what explains take-up rates and different choices between (i) -(iii)?

[17]The households could buy up to two ITNs.

To answer this question, the lab-like field experiment elicited these households' time preferences, slightly modifying the standard lab version of the intertemporal "preference reversal" experiment.[18] The experimenters call this elicitation process a 'baseline survey,' but the technique comes from lab experiments, and also the elicitation of preferences are incentivized. Therefore in our exposition it is more appropriate to call it a lab-in-the-field experiment (LFE).

In this experiment, the households were asked to choose between 12 binary intertemporal payments, e.g., 10 rupee to be paid one month later vs. 15 rupee four months later; 10 rupee four months later vs. 15 rupee seven months later.[19] This design enabled the researchers to (i) identify time-inconsistent choice patterns, such as choosing 10 rupee in the first set while choosing 15 rupee in the second set, and (ii) associate this choice data to the choice data from the RFE concerning the purchase of ITNs. In addition, the experimenters also elicited households' beliefs about the risk of becoming infected with malaria within a year when sleeping regularly under a treated net, an untreated net, and no net in cases where one is a child, a pregnant woman, or an adult. Based on these data and the literature on intertemporal decision-making, the researchers hypothesized three types of households: (1) time-consistent agents; (2) time-inconsistent agents who are aware of their own future present-bias ("sophisticated" time-inconsistent agents); and (3) time-inconsistent agents who are unaware of their own future present-bias ("naive" time-inconsistent agents).

Using these types, the experimenters identified a full structural model of the data-

[18]The experimenters modified some details based on contextual considerations. See Mahajan and Tarozzi (2012, footnote 14, page 7)

[19]The participants were told that one of the 12 chosen rewards would be paid by the local micro-lending association.

generating process underlying the take-up choices in the RFE, and then estimated its parameters using standard goodness of fit methods (maximum likelihood estimation with the finite mixture approach). An important point is that not only was the data-generating process of the RFEs have been fully modelled, but also the model was constructed and estimated based on the LFE, which revealed the heterogeneity of agent types and other characteristics. In other words, the LFE helped solve the so-called identification problem by providing data about the subjects' risk preferences, cost preferences, and beliefs regarding risks of malaria infection. Analysing these data from the LFE, together with the choice data from the RFE in which realistic services (nets with different loan instruments) are offered, enabled the researchers to study which types of subjects chose which micro-credit loans.

One of the main findings of Mahajan and Tarozzi (2012) is that the mapping between the subjects' types (40% of the subjects are consistent, 50% "naive" hyperbolic, and 10% "sophisticated" hyperbolic) and the field purchase patterns is not straightforward. In particular, they found that the product with commitment (i.e., the loan with future re-treatment of the net included) was more popular among wealthier (and "naive") households than among the "sophisticated" ones who, according to the received view, should be the main target of such a product. They also found that the belief in ITN efficacy and wealth predict the choice of the commitment device better than being "sophisticated" does. This result challenges the prominent behavioral economic idea that the poor's under-investment in preventive health products is caused by hyperbolic time discounting.

As this example illustrate, a hybrid design combining LFEs and RFEs can ex-

ploit their complementarities as exhibits. First, a series of prominent RFEs have exhibited the poor's under-investment in preventive health products as a prominent phenomenon to be investigated further (①) in Table 2). Second, an RFE has been designed as an analogue of real-world interventions, in particular different micro credit schemes, *together with* an LFE that measured households' relevant constructs such as risk and cost preferences, as well as beliefs regarding the efficacy of bednets (②) in Table 2). Contingent on the validity of these constructs, these data then help identify the data-generating processes underlying the RFE, in the form of a structural model, which is needed for explicit and quantitative evaluations of analogous counterfactual interventions (③) in Table 2).

Our proposed strategy is a schematic sketch of how research can proceed, so the temporal order of the actual processes can vary. In particular, as we noted earlier, the validity of experimental constructs cannot be verified simply by combining it with RFEs, but instead it has to be gradually developed as the convergence of different experimental and non-experimental measures of a given construct in the long run. Further, the generalizability of the results from a RFE faces more challenges as the intended context of application moves further away from the original RFE, for example to different populations or to a larger scale, which cannot be fully addressed by structural modelling. Despite all these challenges, we suggest that coupling RFEs and LFEs is a concrete and promising step forward to strategically exploit low-hanging methodological advantages of the two strands of field experiments.

5 Conclusion

In this paper, we argued that field experiments in economics come from two distinct historical strands, the first as an extension of laboratory experiments, which we labelled as lab-in-the-field experiments (LFEs), and the second strand defined by the implementation of randomized-control field experiments (RFEs), originating from social sciences more broadly. While the LFEs strand developed to decrease the artificiality that tight laboratory control can create, the RFEs strand tries to achieve control indirectly for both observable and unobservable variables. Thus the two strands face two distinct external validity issues: artificiality for the LFE strand and generalizability for the RFE strand. While both are serious challenges that cannot be fully addressed in this paper, we proposed a promising methodological strategy to couple RFEs and LFEs, building on Sugden's functional analysis of experiments as exhibits for empirical investigations. Although still in an early stage, several hybrid experiments in the same spirit have begun to appear (see Viceisza, 2016, Table 1). Our historical and methodological analysis clarifies why such hybridization is happening—because these strands have been distinct traditions in field experiments in economics—and why it is a good idea—because it can exploit methodological complementarities of both strands.

References

Acemoglu, D. (2010). Theory, general equilibrium, and political economy in development economics. *Journal of Economic Perspectives*, 24(3):17–32.

Akerlof, G. A. (1982). Labor Contracts as Partial Gift Exchange*. *The Quarterly Journal of Economics*, 97(4):543–569.

Andersson, F. N. and Holm, H., editors (2002). *Experimental Economics: Financial Markets, Auctions, and Decision Making: Interviews and Contributions from the 20th Arne Ryde Symposium*. Springer Science & Business Media, New York.

Banerjee, A., Chassang, S., and Snowberg, E. (2017). Decision theoretic approaches to experiment design and external validity. In Banerjee, A. V. and Duflo, E., editors, *Handbook of Field Experiments*, volume 1 of *Handbook of Economic Field Experiments*, chapter 4, pages 141 – 174. North-Holland.

Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics*, 11:122–133.

Bardsley, N., Cubitt, R., Loomes, G., Moffat, P., Starmer, C., and Sugden, R. (2010). *Experimental economics: Rethinking the rules*. Princeton University Press.

Barrett, C. B. and Carter, M. R. (2010). The power and pitfalls of experiments in development economics: Some non-random reflections. *Applied Economic Perspectives and Policy*, 32(4):515–548.

Basu, K. (2014). Randomisation, causality and the role of reasoned intuition. *Oxford Development Studies*, 42(4):455–472.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122 – 142.

Bobonis, G. J., Miguel, E., and Puri-Sharma, C. (2006). Anemia and school participation. *Journal of Human resources*, 41(4):692–721.

Bolton, G. E. and Ockenfels, A. (2012). Behavioral economic engineering. *Journal of Economic Psychology*, 33(3):665–676.

Camerer, C. F. (2003). *Behavioral Game Theory*. Princeton University Press, Princeton, NJ.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological bulletin*, 54(4):297.

Cartwright, N. (2007). Are rcts the gold standard? *BioSocieties*, 2(1):11–20.

Cartwright, N. (2009). What is this thing called "efficacy"? In Mantzavinos, C., editor, *Philosophy of the Social Sciences: Philosophical Theory and Scientific Practice*, chapter 7, pages 185–206. Cambridge University Press, Cambridge, England.

Cartwright, N. (2010). What are randomised controlled trials good for? *Philosophical Studies*, 147(1):59.

Cartwright, N. and Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Charness, G., Gneezy, U., and Kuhn, M. A. (2013). Experimental methods: Extra-laboratory experiments—extending the reach of experimental economics. *Journal of Economic Behavior & Organization*, 91(0):93–100.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Cohen, J. and Dupas, P. (2010). Free distribution or cost-sharing? evidence from a randomized malaria prevention experiment. *The Quarterly Journal of Economics*, 125(1):1–45.

Davis, J. (2013). Economics imperialism under the impact of psychology: The case of behavioral development economics. *Œconomia*, 3(1):119–138.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, 48(2):424–55.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature*, 47(2):315–72.

Duflo, E. (2006). *Field Experiments in Development Economics*, volume 2 of *Econometric Society Monographs*, pages 322–348. Cambridge University Press.

Dupas, P. (2009). What matters (and what does not) in households' decision to invest in malaria prevention? *American Economic Review*, 99(2):224–230.

Dupas, P. (2011). Health behavior in developing countries. *Annual Review of Economics*, 3(1):425–449.

Dupas, P. (2014). Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *Econometrica*, 82(1):197–228.

Fehr, E., Kirchsteiger, G., and Riedl, A. (1993). Does Fairness Prevent Market Clearing? An Experimental Investigation*. *The Quarterly Journal of Economics*, 108(2):437–459.

Ferber, R. and Hirsch, W. Z. (1982). *Social experimentation and economic policy*. Cambridge University Press, Cambridge, England.

Gerber, A. S. and Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. WW Norton.

Gneezy, U., Haruvy, E., and Yafe, H. (2004). The inefficiency of splitting the bill*. *The Economic Journal*, 114(495):265–280.

Gneezy, U. and List, J. A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, 74(5):1365–1384.

Greenberg, D. H. and Shroder, M. (2004). *The digest of social experiments*. The Urban Insitute Press, Washington, D.C., 3rd edition.

Guala, F. (1999). The problem of external validity (or "parallelism") in experimental economics. *Social science information*, 38(4):555–573.

Guala, F. (2003). Experimental localism and external validity. *Philosophy of Science*, 70:1195–1205.

Guala, F. (2005). *The Methodology of Experimental Economics*. Cambridge University Press, Cambridge, England.

Guala, F. (2008a). *Experimental Economics, History of*, pages 1958–1962. Palgrave Macmillan UK, London.

Guala, F. (2008b). Paradigmatic experiments: The ultimatum game from testing to measurement device. *Philosophy of Science*, 75:658–669.

Guala, F. (2010). Extrapolation, analogy, and comparative process tracing. *Philosophy of Science*, 77(5):pp. 1070–1082.

Guala, F. and Mittone, L. (2005). Experiments in economics: External validity and the robustness of phenomena. *Journal of Economic Methodology*, 12(4):495–515.

Guala, F. and Mittone, L. (2010). Paradigmatic experiments: The dictator game. *The Journal of Socio-Economics*, 39(5):578–584.

Guyatt, G., Cairns, J., Churchill, D., Cook, D., Haynes, B., Hirsh, J., Irvine, J., Levine, M., Levine, M., Nishikawa, J., Sackett, D., Brill-Edwards, P., Gerstein, H., Gibson, J., Jaeschke, R., Kerigan, A., Neville, A., Panju, A., Detsky, A., Enkin, M., Frid, P., Gerrity, M., Laupacis, A., Lawrence, V., Menard, J., Moyer, V., Mulrow, C., Links, P., Oxman, A., Sinclair, J., and Tugwell, P. (1992). Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine. *JAMA*, 268(17):2420–2425.

Harrison, G. W. (2011). Randomisation and its discontents. *Journal of African Economies*, 20(4):626–652.

Harrison, G. W. (2013). Field experiments and methodological intolerance. *Journal of Economic Methodology*, 20(2):103–117.

Harrison, G. W. and List, J. A. (2004). Field experiments. *Journal of Economic Literature*, 42(4):1009–1055.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66(5):1017–1098.

Heckman, J. J. (1992). Randomization and social policy evaluation. In Manski, C. F. and Garfinkel, I., editors, *Evaluating Welfare and Training Programs*, chapter 5, pages 201–230. Harvard University Press.

Heckman, J. J., Smith, J., and Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4):487–535.

Heckman, J. J. and Smith, J. A. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives*, 9(2):85–110.

Henrich, J. P. (2004). *Foundations of human sociality: economic experiments and ethnographic evidence from fifteen small-scale societies*. Oxford University Press, Oxford.

Hertwig, R. and Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, 24:383–+.

Heukelom, F. (2009). *Kahneman and Tversky and the Making of Behavioral Economics*. PhD thesis, University of Amsterdam.

Heukelom, F. (2011). What to conclude from psychological experiments: The contrasting cases of experimental and behavioral economics. *History of Political Economy*, 43(4):649–681.

Heukelom, F. (2014). *Behavioral economics: a history*. Cambridge University Press, Cambridge.

Jiménez-Buedo, M. (2011). Conceptual tools for assessing experiments: some well-entrenched confusions regarding the internal/external validity distinction. *Journal of Economic Methodology*, 18(3):271–282.

Jimenez-Buedo, M. and Guala, F. (2016). Artificiality, reactivity, and demand effects in experimental economics. *Philosophy of the Social Sciences*, 46(1):3–23.

Khosrowi, D. (2019). Extrapolation of causal effects – hopes, assumptions, and the extrapolator's circle. *Journal of Economic Methodology*, 26(1):45–58.

Leamer, E. E. (2010). Tantalus on the road to asymptopia. *Journal of Economic Perspectives*, 24(2):31–46.

Ledyard, J. (1995). Public goods: a survey of experimental research. In Kagel, J. H. and Roth, A. E., editors, *The handbook of experimental economics*, pages 111–194. Princeton University Press.

Levitt, S. D. and List, J. A. (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review*, 53(1):1 – 18.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, 115(3):482–493.

List, J. A. and Lucking-Reiley, D. (2000). Demand reduction in multiunit auctions: Evidence from a sportscard field experiment. *American Economic Review*, 90(4):961–972.

List, J. A. and Shogren, J. F. (1998). Calibration of the difference between actual and hypothetical valuations in a field experiment. *Journal of Economic Behavior & Organization*, 37(2):193 – 205.

Mahajan, A. and Tarozzi, A. (2012). Time inconsistency, expectations and technology adoption: The case of insecticide treated nets. WPS 019, University of Californica Berkeley.

Miguel, E. and Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1):159–217.

Ravallion, M. (2009). Should the randomistas rule? *The Economists' Voice*, 6(2).

Reiss, J. (2018). Against external validity. *Synthese*.

Rodrik, D. (2008). The new development economics: We shall experiment, but how shall we learn? *HKS Working Paper No. RWP08-055.* Revised, October 2008.There is a lot more convergence between macro- and micro-development economists than meets the eye.

Ross, D. (2019). Empiricism, sciences, and engineering: cognitive science as a zone of integration. *Cognitive Processing*, pages https://doi.org/10.1007/s10339–019–00916–z.

Roth, A. E. (1995). Introduction to experimental economics. In Kagel, J. H. and Roth, A. E., editors, *Handbook of Experimental Economics*, chapter 1. Princeton University Press.

Schelling, T. C. (1960). *The strategy of conflict*. Harvard University Press.

Shadish, W., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Company, Boston, New York.

Smith, V. L. (1976). Experimental economics: Induced value theory. *The American Economic Review*, 66(2):274–279.

Steel, D. (2008). *Across the boundaries: extrapolation in biology and social science*. Oxford University Press, Oxford.

Steel, D. (2010). A new approach to argument by analogy: Extrapolation and chain graphs. *Philosophy of Science*, 77(5):pp. 1058–1069.

Sugden, R. (2005). Experiments as exhibits and experiments as tests. *Journal of Economic Methodology*, 12(2):291–302.

Sugden, R. (2008). The changing relationship between theory and experiment in economics. *Philosophy of Science*, 75(5):pp. 621–632.

Svorenčík, A. and Maas, H. (2015). *The making of experimental economics: Witness seminar on the emergence of a field*. Springer.

Tarozzi, A., Mahajan, A., Blackburn, B., Kopf, D., Krishnan, L., and Yoong, J. (2014). Micro-loans, insecticide-treated bednets, and malaria: Evidence from a randomized controlled trial in orissa, india. *American Economic Review*, 104(7):1909–41.

Teira, D. (2013). Blinding and the non-interference assumption in medical and social trials. *Philosophy of the Social Sciences*, 43(3):358–372.

Teira, D. and Reiss, J. (2013). *Causality, Impartiality and Evidence-Based Policy*, pages 207–224. Springer Netherlands, Dordrecht.

Viceisza, A. C. G. (2016). Creating a lab in the field: Economics experiments for policymaking. *Journal of Economic Surveys*, 30(5):835–854.